Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Claudius Korzen

Information Retrieval
WS 2017/2018

http://ad-wiki.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 13

Submit until Tuesday, February 6 at **12:00pm (noon)**

Exercise 1 (15 points)

Write a function *pos_tag* that computes the sequence of POS-tags for a given sentence using the Viterbi algorithm, as explained in the lecture. The sentence is given as an array of words, without punctuation (e.g., a comma or a full stop). The transition probabilities and the word distribution for each tag are given in two files on the Wiki. For your convenience, the TIP file contains code to read these files. Note that there are two special POS-tags for the beginning and the end of a sentence.

As usual, you must implement the given test cases: it's easy to make a small mistake and get bad results, which you might not realize without test cases.

Exercise 2 (5 points)

Write a function *find_named_entities* that recognizes entities in a given sentence. As explained in the lecture, you can simply POS-tag the sentence and take each maximal sequence of words tagged with 'NNP' as a named entity.

Write a main function that takes a sentence as input and outputs the sentence with POS tags and the named entities emphasized.

Optionally, disambiguate each named entity to the most popular entity from Wikidata with that name or synonym, using the Wikidata entities file from ES5.

Also optionally, instead of handling the input/output via the console, write a small web app that accepts a sentence in a text box and then outputs the sentence with the POS-tags and entity links. You can use a library for the web server, for example, SimpleHTTPServer for Python. Note that this is not much more work than command-line input/output.

Add your code to a new sub-directory *sheet-13* to our SVN, and make sure that everything runs through without errors on Jenkins. As usual, please briefly tell us about your *experiences* with this exercise sheet. Contemplate on the fact that this is the last exercise sheet.